

Semantische Textanalytik I: Grundlagen des Information Retrieval

Seminar im Modul M-GSW-09
WiSe 2016/17

Prof. Dr. Udo Hahn


Lehrstuhl für Angewandte Germanistische Sprachwissenschaft /
Computerlinguistik

Institut für Germanistische Sprachwissenschaft

Friedrich-Schiller-Universität Jena

<http://www.julielab.de>

Allgemeine Hinweise

- Termin: Do, 16-18h (Humboldt 11, SR 115)
- Materialien im Netz
 - <http://www.julielab.de>  „Students“
- Sprechstunde: Mi, 12-13h (FüG 30, R 203)
- Email: udo.hahn@uni-jena.de
- Fachliteratur: durchgängig in Englisch

Grundlagen des Information Retrieval

Sammeln von Dokumentkollektionen vs. Erschließung von Dokumentinhalten



SRU Sachverständigenrat für Umweltfragen

Der Vorsitzende
Prof. Dr.-Ing. Martin Faustlich
Technische Universität München
Lehrstuhl für Rohstoff- und
Energietechnologie
Patergasse 18
9415 Straubing
Tel. 09421 / 187 100
Fax 09421 / 187 111
martin.faustlich@wzwl.tum.de
www.umweltsr.de

13. Februar 2009

Kommentare des SRU zum Entwurf des Nationalen Biomasseaktionsplans

Sehr geehrte Frau Dr. Freier, sehr geehrter Herr Dr. Ohlhoff,

die Chancen und Risiken der Nutzung nachwachsender Rohstoffe als erneuerbare Energieträger sind in den letzten Jahren intensiv erforcht und diskutiert worden. Angesichts der Komplexität des Themas und der Vielfalt der Nutzungsoptionen ist eine integrierte strategische Betrachtung des Sektors dringend geboten. Vor diesem Hintergrund ist es begrüßenswert, dass die Bundesregierung mit einem nationalen Biomasseaktionsplan den Vorgaben der EU-Biomassestrategie nachkommt. Angesichts der Tatsache, dass sowohl auf der europäischen als auch auf der nationalen Ebene die wesentlichen Ziele und Instrumente bereits rechtlich fixiert sind, kann ein solcher Plan allerdings nur im Detail nachjustieren.

Begrüßenswert ist, dass der Aktionsplan die verstärkte Erzeugung von Wärme, die Erschließung neuer Biomassepotenziale insbesondere aus Reststoffen und Abfällen, die Sicherung der nachhaltigen Erzeugung, die verstärkte Nutzung von Verwertungsoptionen mit besonderem Treibhausgas-Minderungspotenzial, den Vorrang der stofflichen Verwertung, sowie eine verstärkt dezentrale Nutzung im Sinne der Entwicklung der ländlichen Räume als strategische Ziele festschreibt. Bedauerlich ist aber, dass er an vielen Stellen entgegen dem wissenschaftlichen Erkenntnisstand an der grundsätzlichen Gleichwertigkeit der Biomasse-

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN"
"http://www.w3.org/TR/REC-html40/loose.dtd">
<html>
<head>
<title>Welcome to Amaya</title>
<meta name="GENERATOR" content="amaya V2.4">
<meta http-equiv="Content-Type" content="text/html; charset=
">
</head>
<body style="background-color:white" lang="en" text="black">
<h1 style="font-size: 30pt; color: #FF0080">Welcome to Amaya</h1>
<div style="text-align: right; color: #BEBE00">Release 2.4</div>
<strong>Amaya</strong> is a Web client that acts both as an authoring
tool. It has been designed with the primary purpose of supporting
new Web technologies in a WYSIWYG environment on all platforms.
It implements HTML, XHTML, MathML, CSS, and HTTP.</p>
</body>
</html>
```

Süddeutsche Zeitung
NEUESTE NACHRICHTEN AUS POLITIK, KULTUR, WIRTSCHAFT UND SPORT

„Kauft eure Drogen woanders“ – Krawall in Kopenhagen Panorama

SPD stärkste Partei, AfD liegt vor CDU

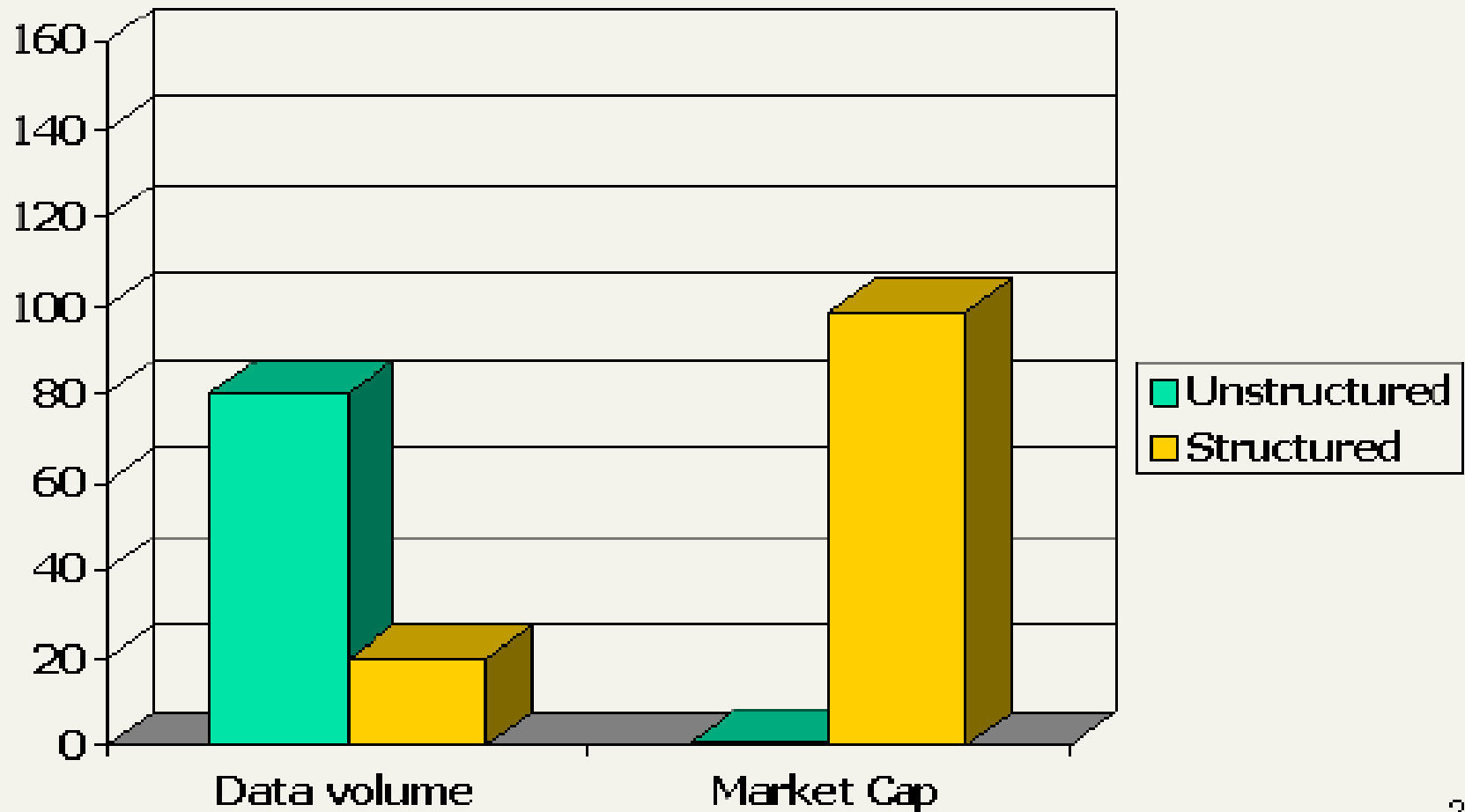
Kaufte eure Drogen woanders?

Krawall in Kopenhagen

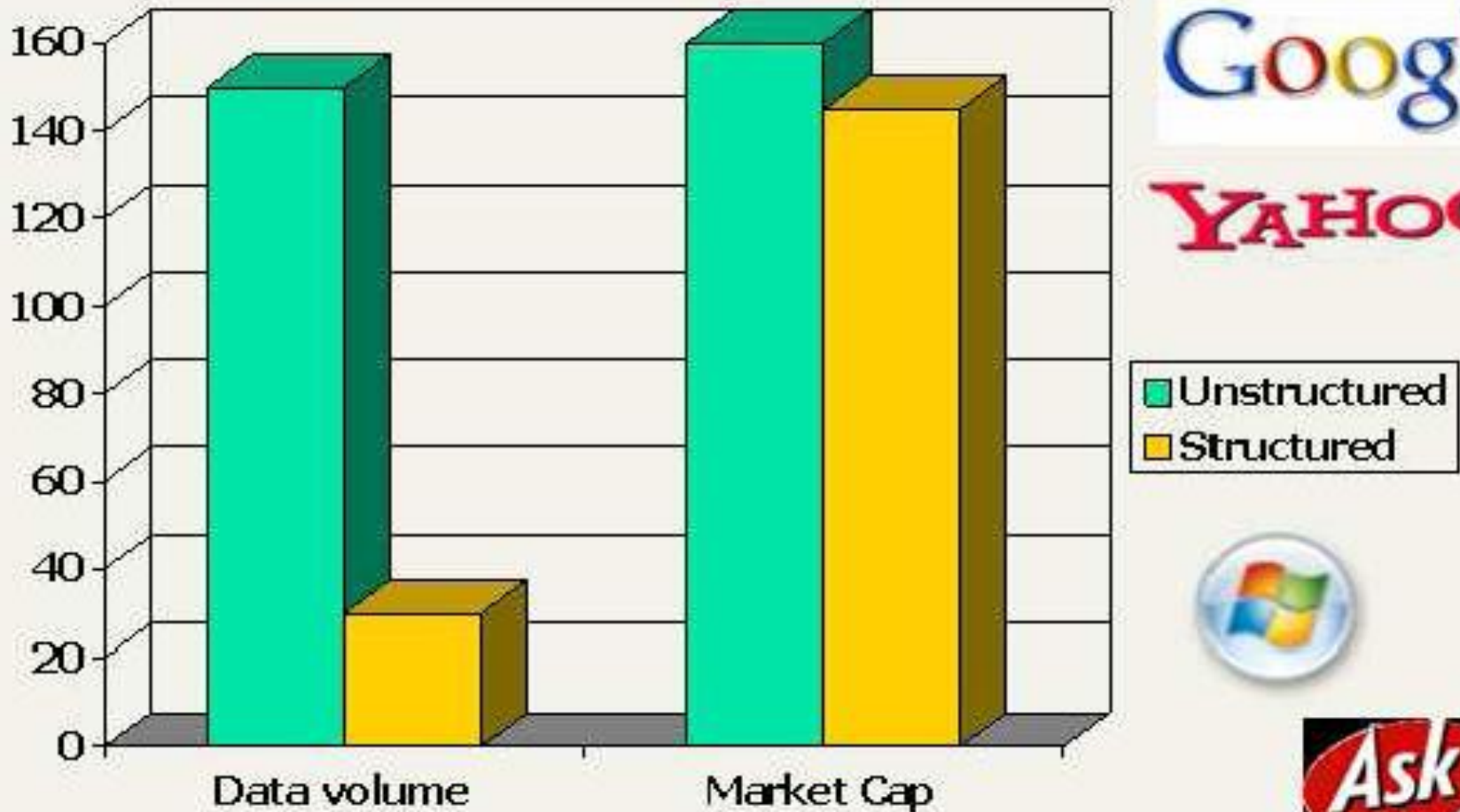
SPD-Präsidium will Zustimmung zu Ceta-Abkommen

lease 2.4
t has been
a fashML,
les and the
nants
cuments on
d edit Web
simple
k, you
er
ra. By this
tion, you
so possible
s

Strukturierte vs. unstrukturierte Daten (1996)



Strukturierte vs. unstrukturierte Daten (2006)



Google™

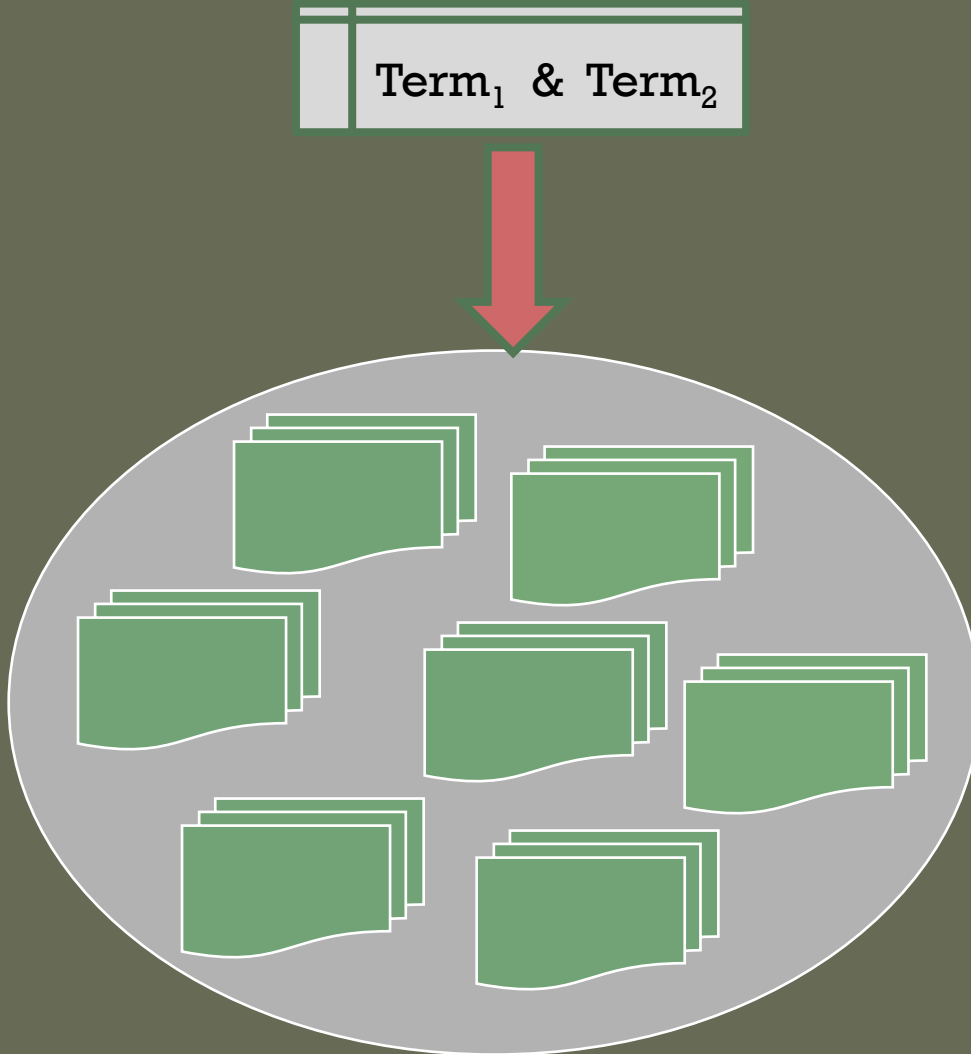
YAHOO!

Unstructured
Structured



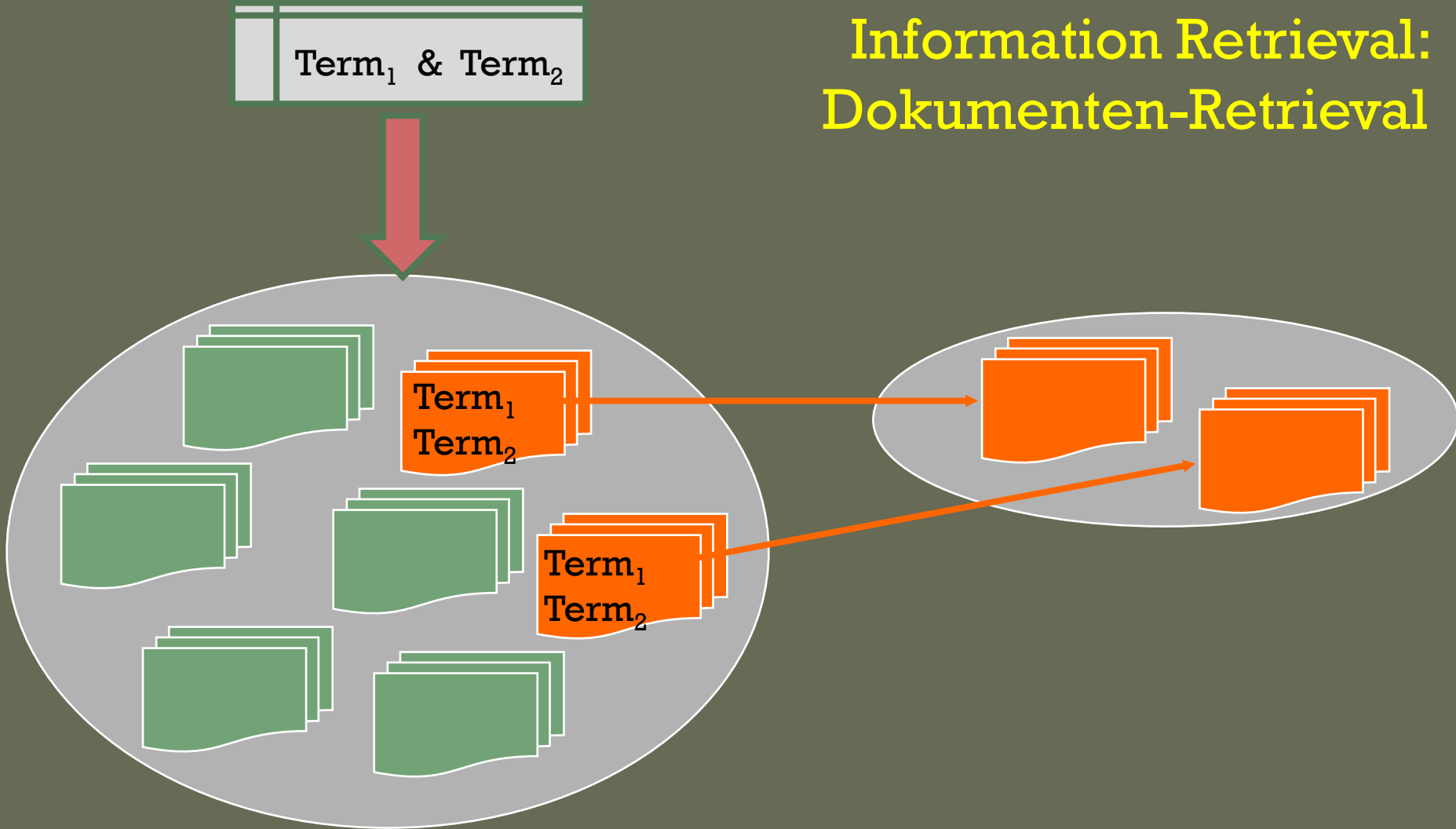
Grundlagen des Information Retrieval

Information Retrieval: Dokumenten-Retrieval



Grundlagen des Information Retrieval

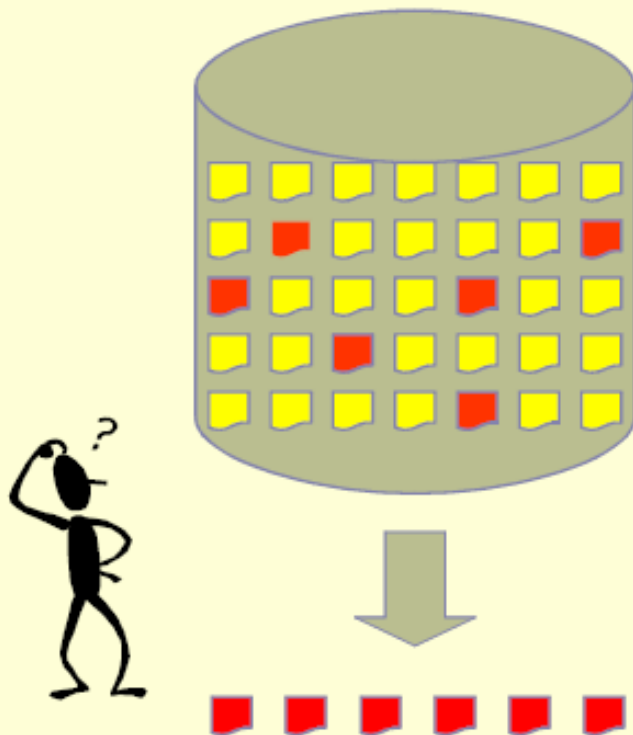
Information Retrieval: Dokumenten-Retrieval



Ad hoc Retrieval vs. Routing

Ad-hoc retrieval

One time queries (e.g. Web search)



Filtering/Routing

Constant search profile (e.g. Spam filtering)



Textkategorisierung: Clustering vs. Themenerkennung

- **Categorization/Clustering:**

Group documents into predefined classes/ adaptive clusters



- **Topic Detection and Tracking:**

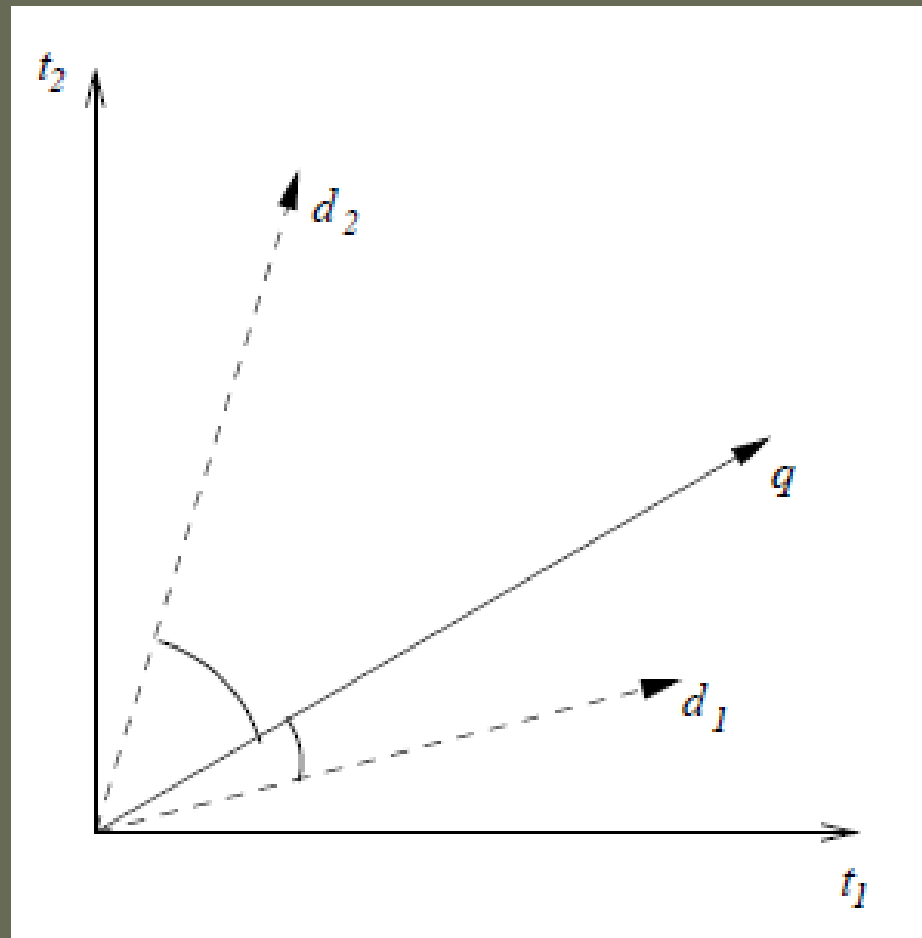
Cluster news in stream

A B D A C D E B D C E B A

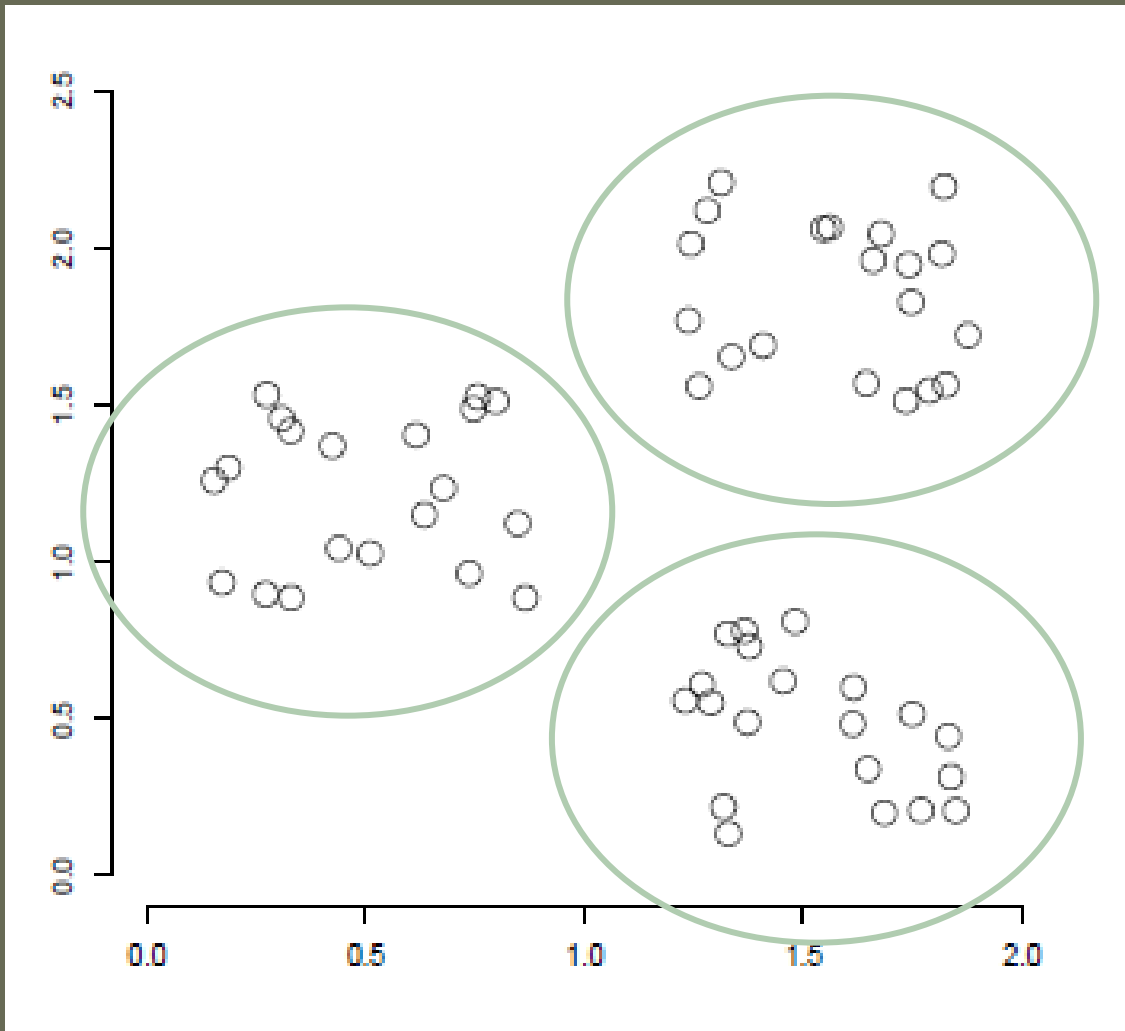
Information Retrieval Modelle

- Klassisches Boole'sches Retrieval-Modell
 - Boole'sche Operatoren (UND, ODER, NICHT)
 - Textstruktur-Operatoren (NEXT, IN-UNIT)
- Vektorraum-Modell
 - Dokument-Term-Matrizen
 - Ähnlichkeitsmaße (Cosinus, Jaccard usw.)
 - Clustering-Verfahren
- Probabilistische Modelle
 - Annahme spezieller Verteilungen von Termen

Vektorraum-Modell



Clustering



Experimentelle Rahmenbedingungen

- ◎ „Relevanz“
- ◎ Retrieval-Metriken
 - Recall, Precision, F-Score
- ◎ Retrieval-Wettbewerbe
 - TREC (seit 1992, jährlich)
 - Dutzende von Tracks: Ad hoc, Routing, QA, video, ...
 - CLEF (seit 2000, jährlich)
 - Multi/cross-lingual, nicht-englische Sprachen

Seminarleistungen

◎ **Vortrag (mündlich)**

- 1-stündig
- Elektronische Version (PDF, PPT) verfügbar machen

◎ **Referat (schriftlich)**

- 15-20 Seiten Kerntext (mit Standardformaten)
- Elektronische Version (PDF, DOC) verfügbar machen
- Eidesstattliche Erklärung zur Eigenautorenschaft
 - Wir prüfen mit Plagiatserkennungs-Software
- Abgabe: Anfang März 2017

Bemerkungen zu Referaten

- **Aufbaumuster:**
 - Deck- bzw. Titelblatt mit vollständigen Angaben
 - Inhaltsverzeichnis
 - Einführung ins Thema, Motivation
 - Themenabhandlung: grundlegende Formalisierungen, Verfahrensbeschreibungen (Algorithmen), Systemfunktionalitäten, Ressourcenmerkmale, Experimente/Evaluationen usw.
 - Fazit mit kritischer Würdigung, offene Probleme ansprechen
 - Bibliographie
- **Zitationen:**
 - Alle verwendeten Quellen zitieren
 - Mit einem bibliographisch korrektem Zitat die jeweilige Quelle eindeutig beschreiben
 - Fachartikel nicht mit **http://...foo.pdf**-Link zitieren
 - Online-Quellen mit URLs und Datum des letztem Zugriffs
 - **Wikipedia** ist keine zitierfähige wissenschaftliche Quelle !
- **Eigenleistungen** (Literatur, Beschäftigung mit konkreten Ressourcen/Systemen usw.) sind sehr erwünscht → unabdingbar !

Wege zum Vortrag und Referat

- Email: Anmeldung von **drei** nach fallender Priorität geordneten Themenwünschen
 - First-come, first-served
- Email: Themenvergabe durch Dozenten
- Erste Literaturhinweise als „Saat“ nach Bestätigung der Themenauswahl
- Themenbearbeitung durch Referenten
 - Mündlicher Vortrag zum vereinbarten Termin
 - Schriftliches Referat (unter Einhaltung der organisatorischen Verabredungen) zum vereinbarten Termin

Grundlegende Literatur

- * Baeza-Yates, Ricardo A., & Berthier Ribeiro-Neto (1999). *Modern Information Retrieval*. New York, N.Y.: ACM Press/Addison-Wesley.
- * Manning, Christopher D., Prabhakar Raghavan, & Schütze, Hinrich (2008). *Introduction to Information Retrieval*. New York/NY, USA: Cambridge University Press
- Büttcher, Stefan, Charles L. A. Clarke, & Gordon V. Cormack (2010). *Information Retrieval: Implementing and Evaluating Search Engines*. Cambridge/MA: MIT Press
- Croft, W. Bruce, D. P. Metzler, & T. Strohman (2009). *Search Engines: Information Retrieval in Practice*. Reading/MA etc.: Addison-Wesley Publishing Company

Wichtige Zeitschriften

- Information Processing & Management
- Journal of the American Society for Information Science
- Journal of Documentation
- Foundations and Trends® in Information Retrieval
- Information Retrieval Journal
- Annual Review of Information Science and Technology (ARIST)

Wichtige Konferenzen

- ◎ **SIGIR – Proceedings of the nth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval**
- ◎ **CIKM – Proceedings of the nth ACM Conference on Information and Knowledge Management**

Ablaufplan

20.10.	Hahn
27.10.	Hahn – Themenvergabe
03.11.	---
10.11.	---
17.11.	---
24.11.	---
01.12.	---
08.12.	---
15.12.	---
22.12.	---
05.01.	Reiter: Grundlagen IR: Boolesch, VR
12.01.	Glaß: Web-Retrieval: PageRank u.a.
19.01.	Ziemer: QA
26.01.	
02.02.	---